Convergence of adaptive sampling schemes

R. Douc, A. Guillin, Jean-Michel Marin, C.P. Robert

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Convergence of adaptive sampling schemes

R. Douc — A. Guillin — J.M. Marin — C.P. Robert

N° 5485

Février 2005

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--5485--FR+ENG

INRIA
FUTURS

Convergence of adaptive sampling schemes

R. Douc* , A. Guillin† , J.M. Marin‡ , C.P. Robert§

Thème COG — Systèmes cognitifs
Projets Select

Rapport de recherche n° 5485 — Février 2005 — 23 pages

Abstract: In the design of efficient simulation algorithms, one is often beset with a poor choice of proposal distributions. Although the performances of a given kernel can clarify how adequate it is for the problem at hand, a permanent on-line modification of kernels causes concerns about the validity of the resulting algorithm. While the issue is quite complex and most often intractable for MCMC algorithms, the equivalent version for importance sampling algorithms can be validated quite precisely. We derive sufficient convergence conditions for a wide class of population Monte Carlo algorithms and show that Rao–Blackwellized versions asymptotically achieve an optimum in terms of a Kullback divergence criterion, while more rudimentary versions simply do not benefit from repeated updating.

Key-words: Central Limit Theorem, importance sampling, Kullback divergence, Law of Large Numbers, MCMC, population Monte Carlo, Rao-Blackwellization.

† CEREMADE, University Paris Dauphine and TSI, ENST, Paris, France
* CMAP, Polytechnique, Palaiseau, France
‡ Project SELECT, INRIA FUTURS, University Paris-Sud, Orsay and CEREMADE, University Paris Dauphine, France
§ CEREMADE, University Paris Dauphine and CREST, INSEE, Paris, France

Unité de recherche INRIA Futurs
Parc Club Orsay Université, ZAC des Vignes,
4, rue Jacques Monod, 91893 ORSAY Cedex (France)
Téléphone : +33 1 72 92 59 00 —Télécopie : +33 1 72 92 59 ??

Convergence de schémas de simulation adaptatifs

Résumé : Dans de nombreux problèmes pratiques, on est amené à calculer une loi de probabilité conditionnelle ou marginale à partir d'une loi jointe. Ce calcul est bien souvent impossible à effectuer explicitement. On peut alors construire un schéma de simulation basé sur l'utilisation de lois instrumentales. Dans ce cas, un mauvais choix lié à l'usage de ces lois peut se réveler désastreux. Dans cet article, nous étudions le comportement théorique d'une classe de schémas d'échantillonnage préférentiel adaptatifs appelés D-kernels Population Monte Carlo. Nous montrons que leur version de Rao-Blackwell s'adapte parfaitement à la loi cible en convergeant vers un optimum au sens de la divergence de Kullback.

Mots-clés : Théorème Central Limite, échantillonnage préférentiel, divergence de Kullback, Loi des Grands Nombres, algorithmes MCMC, schémas Population Monte Carlo.

1 Introduction

In the simulation settings found in optimization and (Bayesian) integration, it is well-documented (Robert and Casella, 2004) that the choice of the instrumental distributions is paramount for the efficiency of the resulting algorithms. Indeed, whether we are considering implementing a Metropolis–Hasting algorithm with proposal density $q(x|y)$ or an importance sampling algorithm with importance function $g(x)$, we are relying on a distribution that is customarily difficult to calibrate, outside a limited range of well-known cases. For instance, a standard result is that the optimal importance density for approximating an integral

$$\mathfrak{I} = \int f(x)\pi(x)dx$$

is $g^\star(x) \propto |f(x)|\,\pi(x)$ (Robert and Casella, 2004, Theorem 3.12), but this formal result is not very informative about the practical choice of g, while a poor choice of g may result in an infinite variance estimator. Similarly, it has been established by Mengersen and Tweedie (1996) that the choice of the transition kernel $q(x|y)$ in the Metropolis–Hastings algorithm is crucial for the resulting convergence speed of the Markov chain.

While the goals of simulating experiments are multifaceted and therefore the efficiency of an algorithm can be evaluated under many different perspectives, a measure of agreement between the target and the proposal distribution can serve as a proxy in many cases: In the nomenclature designed in Andrieu and Robert (2001), examples of such measures are moments, acceptance rates and autocorrelations. An even more robust measure is the Kullback divergence, which is ubiquitous in statistical approximation theory (Csizàr and Tusnàdy, 1984) and which will be used in this paper.

Given the complexity of the original optimization or integration problem (which does require Monte Carlo approximations), it is rarely the case that the optimization of the proposal distribution against an efficiency measure can be achieved in closed form. Even the computation of the efficiency measure for a given proposal is impossible in the majority of cases. For this reason, a number of adaptive schemes have appeared in the recent literature (Robert and Casella, 2004, Section 7.6.3), in order to design better proposals against a given measure of efficiency without resorting to a standard optimization algorithm. For instance, in the MCMC community, sequential changes in the variance of Markov kernels have been proposed in Haario et al. (1999, 2001), while adaptive changes taking advantage of regeneration properties of the kernels have been constructed by Gilks et al. (1998) and Sahu and Zhigljavsky (1998, 2003). In a more general perspective, Andrieu and Robert (2001) develop a two-level stochastic optimization scheme to update parameters of a proposal towards a given integrated efficiency criterion like the acceptance rate (or its difference with a value known to be optimal, see Roberts et al., 1997). As reflected by this general technical report of Andrieu and Robert (2001), the complexity of devising valid adaptive MCMC schemes is however a genuine drawback in their extension, given that the constraints on the inhomogeneous Markov chain that results from this adaptive construction either are difficult to satisfy or result in a fixed proposal after a certain number of iterations.

As stressed in Cappé et al. (2004) (see also Robert and Casella, 2004, Chap. 14), the importance sampling perspective is much more amenable to adaptivity than MCMC, due to its unbiased nature: using sampling importance resampling (Rubin, 1987, 1988), any given sample from an importance distribution g can be transformed in a sample of points marginally distributed from the target distribution π and Cappé et al. (2004) showed that this property is also preserved by repeated and adaptive sampling. The asymptotics of adaptive importance sampling are therefore much more manageable than those of adaptive MCMC algorithms, at least at a primary level, if only because the algorithm can be stopped at any time since it does not require a burn-in time. (We will present in this paper more advanced convergence results.) Borrowing from the sequential sampling literature (Doucet et al., 2001), Cappé et al. (2004) constructed an iterative adaptive scheme christened *population Monte Carlo* (Iba, 2000) that aims at replicating the adaptivity of MCMC kernels by a learning mechanism on a population of points, themselves marginally distributed from the target distribution.

In this paper, we establish a CLT for a general PMC scheme and derive an iterative adaptive method that converges to the optimal proposal, the optimality being defined here in terms of Kullback–Leibler divergence. From a probabilistic point of view, the techniques used in this paper are related to techniques and results found in Chopin (2004), Künsch (2004) and Cappé et al. (2005). In particular, the triangular array technique that is central to the CLT proofs below can be found in Cappé et al. (2005) or Douc and Moulines (2005).

The paper is organized as follows: We first present the algorithmic and mathematical details in Section 2. We evaluate the convergence properties of the basic version of PMC in Section 3, exhibiting its limitations, and show in Section 5 that its Rao-Blackwellized version overcomes these limitations and achieve optimality for the Kullback-Leibler criterion developed in Section 4. Section 6 illustrates the practical convergence of the method on a few benchmark examples.

2 Population Monte Carlo

The form of Population Monte Carlo (PMC) introduced in Cappé et al. (2004) intrinsically is a form of iterated sampling importance resampling (SIR), following the device of Rubin (1987, 1988). The idea of using a repeated form of SIR is that previous samples are informative about the connections between the proposal (importance) and the target distributions. We stress from the start that there are very few connections with MCMC algorithms in this scheme since (a) PMC is not Markovian, being possibly based on the whole sequence of simulation, and (b) PMC can be stopped at any time, being validated by the basic importance sampling identity (Robert and Casella, 2004, equation (3.9)) rather than by a probabilistic convergence result like the ergodic theorem. These features motivate the use of the method in setups where off-the-shelf MCMC algorithms cannot be of help. We first recall some basic Monte Carlo techniques to define notations and goals.

2.1 The Monte Carlo framework

On a measurable space $(\Omega, \mathcal{A})$, we consider a probability distribution π on $(\Omega, \mathcal{A})$. We assume that π is dominated by a reference measure μ, $\pi \ll \mu$, and also denote $\pi(dx) = \pi(x)\mu(dx)$ its density. We also suppose that π is known up to a normalizing constant,

$$\pi(x) = \frac{\tilde{\pi}(x)}{\int \tilde{\pi}(x)\mu(dx)}\,,$$

where $\tilde{\pi}$ is known, but the calculation of $\int \tilde{\pi}(x)\mu(dx) < \infty$ is intractable.

For one or several π-measurable functions f, we are interested in computing an approximation of

$$\pi(f) = \int f(x)\pi(dx) = \frac{\int f(x)\tilde{\pi}(x)\mu(dx)}{\int \tilde{\pi}(x)\mu(dx)}\,,$$

assuming that the calculation of $\int f(x)\tilde{\pi}(x)\mu(dx)$ is also intractable.

In this setting, a standard stochastic approximation method is the Monte Carlo method, based on an iid sample $x_1, \ldots, x_N$ simulated from π, that approximates $\pi(f)$ by

$$\hat{\pi}_N^{MC}(f) = N^{-1}\sum_{i=1}^{N} f(x_i)\,,$$

which almost surely converges to $\pi(f)$ (as N goes to infinity) by the Law of Large Numbers (LLN). The Central Limit Theorem (CLT) implies in addition that, if $\pi(f^2) = \int f^2(x)\pi(dx) < \infty$,

$$\sqrt{N}\left\{\hat{\pi}_N^{MC}(f) - \pi(f)\right\} \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}\left(0, \mathbb{V}_\pi(f)\right)\,,$$

where $\mathbb{V}_\pi(f) = \pi([f - \pi(f)]^2)$. Obviously, this approach requires a direct iid simulation from π (or $\tilde{\pi}$) which often is impossible. An alternative (see, e.g., Robert and Casella, 2004, Chap. 3) is to use importance sampling, that is, to pick a probability distribution $\nu \ll \mu$ on $(\Omega, \mathcal{A})$ called the proposal or importance distribution, with density also denoted by ν, and to estimate $\pi(f)$ by

$$\hat{\pi}_{\nu,N}^{IS}(f) = N^{-1}\sum_{i=1}^{N} f(x_i)\left(\frac{\pi}{\nu}\right)(x_i).$$

If π is also dominated by ν, $\pi \ll \nu$, $\hat{\pi}_{\nu,N}^{IS}(f)$ almost surely converges to $\pi(f)$. Moreover, if $\nu\left(f^2\left(\pi/\nu\right)^2\right) < \infty$, the CLT also applies, that is,

$$\sqrt{N}\left\{\hat{\pi}_{\nu,N}^{IS}(f) - \pi(f)\right\} \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}\left(0, \mathbb{V}_\nu\left(f\frac{\pi}{\nu}\right)\right)\,.$$

As the normalizing constant of the target distribution π is unknown, it is not possible to use directly the IS estimator $\hat{\pi}^{IS}_{\nu,N}(f)$ and we need to replace it with the self-normalized version of the IS estimator,

$$\hat{\pi}^{SNIS}_{\nu,N}(f) = \left(\sum_{i=1}^{N} (\pi/\nu)(x_i)\right)^{-1} \sum_{i=1}^{N} f(x_i)\,(\pi/\nu)(x_i)\,,$$

which also converges almost surely to $\pi(f)$. If $\nu\left((1+f^2)\,(\pi/\nu)^2\right) < \infty$, the CLT applies:

$$\sqrt{N}\,\left\{\hat{\pi}^{SNIS}_{\nu,N}(f) - \pi(f)\right\} \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}\left(0, \mathbb{V}_\nu\left\{[f-\pi(f)]\,\frac{\pi}{\nu}\right\}\right)\,.$$

Obviously, the quality of the IS and of the SNIS approximations strongly depends on the choice of the proposal distribution ν, which is delicate for complex distributions like those that occur in high dimensional problems.

2.2 Sampling Importance Resampling

The Sampling Importance Resampling (SIR) method (Rubin, 1987, 1988) is an extension of the IS method that achieves simulation from π by resampling rather than by simple reweighting. More precisely, the SIR algorithm is held in two stages: The first stage is similar to IS and consists in generating an iid sample $x_1, \ldots, x_N$ from ν. The second stage builds a sample from π, $\tilde{x}_1, \ldots, \tilde{x}_M$, based on the instrumental sample $x_1, \ldots, x_N$ by resampling. While there are many resampling methods (Robert and Casella, 2004, Section 14.3.5), the most standard (if least efficient) approach is multinomial resampling in $x_1, \ldots, x_N$ with probabilities proportional to the importance weights $\left[\frac{\pi}{\nu}(x_1), \ldots, \frac{\pi}{\nu}(x_N)\right]$:

$$\tilde{x}_i = x_{J_i}\,, \qquad 1 \le i \le M\,,$$

where the random variables $(J_1, \ldots, J_M)$ are iid conditionally on $x_1, \ldots, x_N$ and distributed as

$$\mathbb{P}\left[J_l = i | x_1, \ldots, x_N\right] = \left(\sum_{j=1}^{N} \frac{\pi}{\nu}(x_j)\right)^{-1} \frac{\pi}{\nu}(x_i)\,.$$

The SIR estimator of $\pi(f)$ is then

$$\hat{\pi}^{SIR}_{\nu,N,M}(f) = M^{-1} \sum_{i=1}^{M} f(\tilde{x}_i)$$

which also converges to $\pi(f)$ since each $\tilde{x}_i$ is (marginally) approximatively distributed from π. By construction, the variance of $\hat{\pi}^{SIR}_{\nu,N,M}(f)$ is larger than the variance of the SNIS estimator. Indeed, the expectation of $\hat{\pi}^{SIR}_{\nu,N,M}(f)$ conditional on the sample $x_1, \ldots, x_N$ is equal to $\hat{\pi}^{SNIS}_{\nu,N}(f)$. Note that an asymptotic analysis of $\hat{\pi}^{SIR}_{\nu,N,M}(f)$ is quite delicate because of the dependencies in the SIR sample (which, again, is not an iid sample from π).

2.3 The Population Monte Carlo algorithm

In their alternative generalization of Importance Sampling, Cappé et al. (2004) introduce an iterative feature in the production of importance samples, for the purpose of adapting the importance distribution ν to the target distribution π. Iterations are indeed necessary to learn about π from the (poor or good) performances of earlier proposals, performances that are for instance evaluated through the distribution of the importance weights. At iteration t of the PMC algorithm, N realizations are thus simulated from a proposal distribution that is derived from the $N \times (t-1)$ previous realizations. Cappé et al. (2004) show that the dependence on earlier proposals and realizations does not jeopardize the fundamental importance sampling identity. Local and adaptive importance sampling schemes can thus be chosen in a much wider generality than thought previously. By introducing a temporal dimension to the selection of the importance function, an adaptive perspective can be achieved at little cost, for a potentially large gain in efficiency.

If we introduce the σ-algebras related to the current and past simulations,

$$\mathcal{F}_{N,t} = \sigma\left\{(x_{i,j}, J_{i,j})_{1\le i\le N, 0\le j\le t}\right\} \quad (t \ge 0)\,,$$
$$\mathcal{F}^J_{N,t} = \mathcal{F}_{N,t} \bigvee \sigma\left\{((x_{i,t+1})_{1\le i\le N}\right\}(t \ge 0)\,,$$

where both the $x_{i,j}$'s and the $J_{i,j}$'s are defined precisely below, and if we set the renormalized importance weights as

$$\bar{\omega}_{i,t} = \frac{\omega_{i,t}}{\sum_{j=1}^{N} \omega_{j,t}},$$

the generic PMC algorithm reads as follows:

Generic PMC algorithm:
At time 0,

a) Generate $(x_{i,0})_{1\leq i\leq N}$ iid according to ν_0 and compute the importance weights $\omega_{i,0} = \{\pi/\nu_0\}(x_{i,0})$;

b) Conditionally on $\sigma\{(x_{i,0})_{1\leq i\leq N}\}$, draw

$$(J_{i,0})_{1\leq i\leq N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\bar{\omega}_{i,0})_{1\leq i\leq N})$$

and set $\tilde{x}_{i,0} = x_{J_{i,0},0}$ $(1 \leq i \leq N)$.

At time t $(t = 1, \ldots, T)$

a) Conditionally on $\mathcal{F}_{N,t-1}$, draw independently $x_{i,t}$ according to $\nu_{i,t}(\mathcal{F}_{N,t-1})$ and compute the importance weights $\omega_{i,t} = \{\pi/\nu_{i,t}(\mathcal{F}_{N,t-1})\}(x_{i,t})$;

b) Conditionally on $\mathcal{F}^J_{N,t-1}$, draw

$$(J_{i,t})_{1\leq i\leq N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\bar{\omega}_{i,t})_{1\leq i\leq N})$$

and set $\tilde{x}_{i,t} = x_{J_{i,t},t}$ $(1 \leq i \leq N)$.

After T iterations of the previous algorithm, a PMC estimator of $\pi(f)$ is given by

$$\hat{\pi}^{PMC}_{N,T}(f) = \left(\sum_{j=1}^{N} \left\{\frac{\pi}{\nu_{i,T}(\mathcal{F}_{N,T-1})}\right\}(x_{j,T})\right)^{-1} \sum_{i=1}^{N} \left\{\frac{\pi}{\nu_{i,T}(\mathcal{F}_{N,T-1})}\right\}(x_{i,T}) f(x_{i,T}),$$

although it is more efficient for all estimation purposes to average the PMC approximations over all iterations, possibly with different weights. Note that we adopt the representation $\nu_{i,t}(\mathcal{F}_{N,t-1})$ for the importance function to signify that the construction of the proposal distribution for the i-th term of the t-th sample is completely open, as illustrated in Cappé et al. (2004). Obviously, all adaptive schemes do not lead to an automatic improvement of the proposal and we now consider two particular schemes where improvement does not occur and does occur, respectively.

3 The D-kernel PMC algorithm

In this section, we introduce a particular PMC scheme for which $\nu_{i,t}(\mathcal{F}_{N,t-1})$ is a mixture of D different transition kernels Q_k $(1 \leq k \leq D)$ that are chosen prior to the simulation experiment, but whose weights are proportional to their survival rates in the previous resampling step. This scheme was first proposed in Cappé et al. (2004), with the purpose that, over iterations, the algorithm would automatically adapt the mixture to the target distribution by converging to the "right" weights, in a spirit similar to the mixture adaption found in Andrieu and Robert (2001). We will however see in this Section that this is not the case, and, more dramatically, that this scheme is intrinsically non-adaptive.

3.1 The algorithm

We consider a family $(Q_d)_{1\leq d\leq D}$ of D transition kernels on on $\Omega \times \mathcal{A}$ and we assume that both π and $(Q_d(x,\cdot))_{1\leq d\leq D,\ x\in\Omega}$ are dominated by the reference measure μ introduced earlier. As above, we also set the corresponding density function and transition kernel to be π and $q_d(\cdot,\cdot)$ respectively, that is

$$\forall A \in \mathcal{A}, \quad \pi(A) = \int_A \pi(x)\mu(dx), \quad Q_d(x,A) = \int_A q_d(x,x')\mu(dx').$$

This situation is rather common in MCMC settings where several vintage transition kernels are often available and difficult to compare. For instance, the cycle and mixture MCMC schemes already discussed by Tierney (1994) are of this nature. We detail in this Section and the following ones how PMC can overcome the difficulty encountered by MCMC algorithms in picking an efficient mixture of standard kernels $\sum_d \alpha_d Q_d(x,\cdot)$.

The associated PMC algorithm then builds proposals as follows:

D-kernel PMC algorithm:
At time 0, use the same step as in the generic PMC algorithm to produce the sample $(\tilde{x}_{i,0}, J_{i,0})_{1\le i\le N}$ and set $\alpha_d^{1,N} = 1/D$ for all $1 \le d \le D$.
At time t ($t = 1, \ldots, T$),

a) Conditionally on $\sigma\left\{(\alpha_d^{t,N})_{1\le d\le D}\right\}$, generate

$$(K_{i,t})_{1\le i\le N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\alpha_d^{t,N})_{1\le d\le D})$$

b) Conditionally on $\sigma\{(\tilde{x}_{i,t-1}, K_{i,t})_{1\le i\le N}\}$, generate independent

$$(x_{i,t})_{1\le i\le N} \sim Q_{K_{i,t}}(\tilde{x}_{i,t-1}, \cdot)$$

and set $\omega_{i,t} = \pi(x_{i,t})/q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})$;

c) Conditionally on $\sigma\{(\tilde{x}_{i,t-1}, K_{i,t}, x_{i,t})_{1\le i\le N}\}$, generate

$$(J_{i,t})_{1\le i\le N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\bar{\omega}_{i,t})_{1\le i\le N})$$

and set ($1 \le i \le N$, $1 \le d \le D$)

$$\tilde{x}_{i,t} = x_{J_{i,t},t}\,, \quad \alpha_d^{t+1,N} = \sum_{i=1}^N \bar{\omega}_{i,t} \mathbb{I}_d(K_{i,t})$$

Recall that $\bar{\omega}_{i,t}$ denotes the renormalized version of $\omega_{i,t}$. In words, Step a) picks the kernel index in the mixture for each point in the sample, Step b) generates the corresponding point and Step c) updates the weights of the D kernels according to their respective survival performances in the past round. (Since survival is directed by the importance weights, reweighting is thus related to the respective magnitudes of the importance weights for the different kernels.) Note also that Step c) is only used to avoid the dissemination of small importance weights along iterations and the subsequent degeneracy phenomenon that plagues iterated IS schemes like particle filters. Integral approximations should however use the byproduct of Step b).

3.2 Convergence properties

In order to assess the average effect of these iterations, we now consider the convergence of the algorithm when the number N of points in each sample is large. Indeed, as already pointed out in Cappé et al. (2004), it does not make much sense to consider the asymptotics of the PMC scheme when T grows large, given that this algorithm is intended to be run with a small number T of iterations.

In order to prove convergence of the D kernel PMC algorithm, we first assume that the generalized importance weight is almost surely finite, that is,

$$\textbf{(A1)} \qquad \forall d \in \{1, \ldots, D\},\ \pi \otimes \pi\{q_d(x,x') = 0\} = 0.$$

Note that assumption **(A1)** implies that $\pi \otimes \pi\left\{\pi(x')/q_d(x,x') < \infty\right\} = 1$. We denote by γ_u the uniform distribution on $\{1, \ldots, D\}$, that is, $\gamma_u(k) = 1/D$ for all $k \in \{1, \ldots, D\}$. We can then deduce a LLN on the pairs $(x_{i,t}, K_{i,t})$ produced by the above algorithm:

Proposition 3.1. *Under **(A1)**, for any function h in $L^1_{\pi\otimes\gamma_u}$ and for all $t \ge 1$,*

$$\sum_{i=1}^N \bar{\omega}_{i,t} h(x_{i,t}, K_{i,t}) \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} \pi \otimes \gamma_u(h).$$

Proof. We proceed by induction wrt t. Using Theorem A.1, the case $t = 1$ is straightforward since this is a direct consequence of the convergence of the importance sampling algorithm. Now, let $t > 1$ and assume that the LLN holds for $t-1$. For $h \in L^1_{\pi\otimes\gamma_u}$, to prove that $\sum_{i=1}^N \bar{\omega}_{i,t} h(x_{i,t}, K_{i,t})$ converges in probability to $\pi\otimes\gamma_u(h)$, we just need to check that

$$N^{-1}\sum_{i=1}^N \frac{\pi(x_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})} h(x_{i,t}, K_{i,t}) \xrightarrow{N\to\infty}_{\mathbb{P}} \pi\otimes\gamma_u(h)\,,$$
$$N^{-1}\sum_{i=1}^N \frac{\pi(x_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})} \xrightarrow{N\to\infty}_{\mathbb{P}} 1\,,$$

where the second convergence is obviously a special case of the first one (with $h = 1$). For the first convergence, applying Theorem A.1 with

$$\begin{cases} U_{N,i} = N^{-1}\frac{\pi(x_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})} h(x_{i,t}, K_{i,t}), \\ \mathcal{G}_N = \sigma\left\{(\tilde{x}_{i,t-1})_{1\le i\le N}, (\alpha_d^{t,N})_{1\le d\le D}\right\}, \end{cases}$$

we only need to check condition (iii). For all $C > 0$, we have

$$\begin{aligned} N^{-1}\sum_{i=1}^N \mathbb{E}\left[\frac{\pi(x_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})} h(x_{i,t}, K_{i,t}) \mathbb{I}_{\left\{\frac{\pi(x_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})} h(x_{i,t}, K_{i,t}) > C\right\}} \middle| \mathcal{G}_N\right] \\ = \sum_{d=1}^D N^{-1}\sum_{i=1}^N F_C(\tilde{x}_{i,t-1}, d)\alpha_d^{t,N} \end{aligned} \tag{1}$$

where $F_C(x,k) = \int \pi(du) h(u,k) \mathbb{I}_{\left\{\frac{\pi(u)}{q_k(x,u)} h(u,k) \ge C\right\}}$. By induction, we have

$$\alpha_d^{t,N} = \sum_{i=1}^N \bar{\omega}_{i,t-1} \mathbb{I}_d(K_{i,t-1}) \longrightarrow_{\mathbb{P}} 1/D$$
$$N^{-1}\sum_{i=1}^N F_C(\tilde{x}_{i,t-1}, k) \xrightarrow{N\to\infty}_{\mathbb{P}} \pi(F_C(\cdot, k))$$

Using these limits in (1) yields

$$N^{-1}\sum_{i=1}^N \mathbb{E}\left[\frac{\pi(x_{i,t}) h(x_{i,t}, K_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})} \mathbb{I}_{\left\{\frac{\pi(x_{i,t}) h(x_{i,t}, K_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})} > C\right\}} \middle| \mathcal{G}_N\right] \xrightarrow{N\to\infty}_{\mathbb{P}} \pi\otimes\gamma_u(F_C).$$

Since $\pi\otimes\gamma_u(F_C)$ converges to 0 as C goes to infinity, this proves that for all $\eta > 0$,

$$N^{-1}\sum_{i=1}^N \mathbb{E}\left(\frac{\pi(x_{i,t}) h(x_{i,t}, K_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})} \mathbb{I}_{\left\{\frac{\pi(x_{i,t}) h(x_{i,t}, K_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})} > N\eta\right\}} \middle| \mathcal{G}_N\right) \xrightarrow{N\to\infty}_{\mathbb{P}} 0\,.$$

Condition (iii) is satisfied and Theorem A.1 applies. The proof follows. □

Note that this convergence result is more than what we need for Monte Carlo purposes since the $K_{i,t}$'s are auxiliary parameters that are not relevant for the original problem. However, it is eventually a negative result in that, while it implies that

$$\sum_{i=1}^N \bar{\omega}_{i,t} f(x_{i,t})$$

is a convergent estimator of $\pi(f)$, it also shows that, for $t \ge 1$,

$$\sum_{i=1}^N \bar{\omega}_{i,t} \mathbb{I}_d(K_{i,t}) \xrightarrow{N\to\infty}_{\mathbb{P}} \frac{1}{D}.$$

Therefore, at *each* iteration, the weights of *all* kernels converge to $1/D$ when the number of points in the sample grows to infinity. This translates in the lack of learning properties for the D-kernel PMC algorithm: its properties at iteration 1 and at iteration 10 are the same. In other words, this algorithm is not adaptive and only requires one iteration with a large value of N. We can note that, when this scheme was used in Cappé et al. (2004), the fast stabilization of the approximation was noticeable. (It is also possible to establish a CLT for this algorithm, but given its unappealing features, we leave the details for the interested reader.)

In order to get a truly adaptive PMC scheme, based on the above D-kernel algorithm, we have first to set an effective criterion of adaptivity and approximation of the target distribution by the proposal distribution. We then derive a modification of the original D-kernel algorithm that achieves efficiency in this sense. As argued in many papers using a wide range of arguments, a natural choice of approximation metric is the Kullback divergence: we can aim at deriving the D-kernel mixture that minimizes the Kullback divergence between this mixture and the target measure π

$$\iint \log\left(\frac{\pi(x)\pi(x')}{\pi(x)\sum_{d=1}^{D}\alpha_d q_d(x,x')}\right)(\pi\otimes\pi)(dx,dx')\,. \tag{2}$$

The following Section is devoted to the problem of finding an iterative choice of mixing coefficients that converges to this minimum. The optimal PMC scheme then follows in Section 5.

4 The Kullback divergence

4.1 The criterion

Using the same notations as above, in conjunction with the choice of the weights α_d in the D kernel mixture, we introduce the simplex of $\mathbb{R}^D$,

$$\mathscr{S} = \left\{\alpha=(\alpha_1,\ldots,\alpha_D);\ \forall d\in\{1,\ldots,D\},\ \alpha_d\geq 0 \quad \text{and} \quad \sum_{d=1}^{D}\alpha_d = 1\right\}$$

and $\bar{\pi}=\pi\otimes\pi$. We then assume that the family of the D kernels satisfies the condition

$$(\mathbf{A2})\ \forall i\in\{1,\ldots,D\},\ \mathbb{E}_{\bar{\pi}}\left[|\log q_i(X,X')|\right] = \iint |\log q_i(x,x')|\bar{\pi}(dx,dx') < \infty,$$

which is automatically satisfied when all q_j's dominate π (in the accept-reject sense that π/q_j is bounded). We then derive from the Kullback divergence a function on $\mathscr{S}$, that is, for $\alpha\in\mathscr{S}$,

$$\mathcal{E}_{\bar{\pi}}(\alpha) = \iint \bar{\pi}(dx,dx')\log\left(\sum_{d=1}^{D}\alpha_d q_d(x,x')\right) = \mathbb{E}_{\bar{\pi}}\left[\log\sum_{d=1}^{D}\alpha_d q_d(X,X')\right].$$

Note that, due to the strict concavity of the log function, $\mathcal{E}_{\bar{\pi}}$ is a strictly concave function on a connected compact set and thus has no local maximum besides the global maximum, denoted

$$\alpha^{max} = \arg\max_{\alpha\in\mathscr{S}} \mathcal{E}_{\bar{\pi}}(\alpha)\,.$$

Note also that, since

$$\int \pi(dx)\log\pi(x) - \mathcal{E}_{\bar{\pi}}(\alpha) = \mathbb{E}_{\bar{\pi}}\left(\log\frac{\pi(X)\pi(X')}{\pi(X)\left\{\sum_{d=1}^{D}\alpha_d q_d(X,X')\right\}}\right),$$

α^{max} is the optimal choice for a mixture of transition kernels such that the joint law of (X_0,X_1) when $X_0\sim\pi$ is the nearest to the product distribution $\bar{\pi}=\pi\otimes\pi$. We then have the following obvious inequality:

Lemma 4.1. *Under **(A1-A2)**, for all $\alpha\in\mathscr{S}$, $\mathcal{E}_{\bar{\pi}}(\alpha)\leq\int\pi(dx)\log\pi(x)$.*

4.2 A maximization algorithm

We now propose an iterative procedure, akin to the EM algorithm, that updates the weights so that the function $\mathcal{E}_{\bar{\pi}}(\alpha)$ increases at each step. We first define F as the function on $\mathscr{S}$ such that

$$F(\alpha) = \left(\mathbb{E}_{\bar{\pi}} \left[\frac{\alpha_d q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \right] \right)_{1 \le d \le D}$$

and construct the sequence on $\mathscr{S}$

$$\begin{cases} \alpha^1 = (1/D, \ldots, 1/D) \\ \alpha^{t+1} = F(\alpha^t) \qquad \text{for} \quad t \ge 1 \end{cases} \tag{3}$$

Note that, under assumption **(A1)**, for all $t \ge 0$,
$\mathbb{E}_{\bar{\pi}} \left(q_d(X, X') \Big/ \sum_{j=1}^D \alpha_j^t q_j(X, X') \right) > 0$ and thus, for all $t \ge 0$ and all $d \in \{1, \ldots, D\}$, $\alpha_d^t > 0$. If we define the extremal set $\mathcal{I}_D$ as

$$\left\{ \alpha \in \mathscr{S}; \forall d \in \{1, \ldots, D\}, \ \alpha_d = 0 \quad \text{or} \quad \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \right) = 1 \right\}, \tag{4}$$

we then have the following fixed point result:

Proposition 4.1. *Under **(A1)** and **(A2)**,*

i) *$\mathcal{E}_{\bar{\pi}} \circ F - \mathcal{E}_{\bar{\pi}}$ is continuous,*

ii) *For all $\alpha \in \mathscr{S}$, $\mathcal{E}_{\bar{\pi}} \circ F(\alpha) \ge \mathcal{E}_{\bar{\pi}}(\alpha)$,*

iii) *$\mathcal{I}_D = \{\alpha \in \mathscr{S}; F(\alpha) = \alpha\} = \{\alpha \in \mathscr{S}; \mathcal{E}_{\bar{\pi}} \circ F(\alpha) = \mathcal{E}_{\bar{\pi}}(\alpha)\}$ and $\mathcal{I}_D$ is finite.*

Proof. $\mathcal{E}_{\bar{\pi}}$ is clearly continuous. Moreover, by Lebesgue dominated convergence theorem, the function $\alpha \mapsto \mathbb{E}_{\bar{\pi}} \left(\alpha_d q_d(X, X') / \sum_{j=1}^d \alpha_j q_j(X, X') \right)$ is also continuous, which implies that F is continuous. This completes the proof of **i)**. Now, by the concavity of the log function,

$$\begin{aligned} &\mathcal{E}_{\bar{\pi}}(F(\alpha)) - \mathcal{E}_{\bar{\pi}}(\alpha) \\ &\quad = \mathbb{E}_{\bar{\pi}} \left(\log \left[\sum_{d=1}^D \frac{\alpha_d q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \right) \right] \right) \\ &\quad \ge \mathbb{E}_{\bar{\pi}} \left[\sum_{d=1}^D \frac{\alpha_d q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \log \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \right) \right] \\ &\quad = \sum_{d=1}^D \alpha_d \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \right) \log \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \right) \end{aligned} \tag{5}$$

Applying the inequality $u \log u \ge u - 1$ to (5) yields **ii)**. Moreover, equality in $u \log u \ge u - 1$ holds if, and only if $u = 1$. Therefore, equality in (5) is equivalent to

$$\forall \alpha_d \ne 0, \quad \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \right) = 1.$$

Thus, $\mathcal{I}_D = \{\alpha \in \mathscr{S}; \mathcal{E}_{\bar{\pi}} \circ F(\alpha) = \mathcal{E}_{\bar{\pi}}(\alpha)\}$. The second equality $\mathcal{I}_D = \{\alpha \in \mathscr{S}; F(\alpha) = \alpha\}$ is straightforward.

We now prove par recursion on D that $\mathcal{I}_D$ is finite. Recalling the definition of $\mathcal{I}_D$, the recursion is quite straightforward. We just need to prove that the set

$$\{\alpha \in \mathcal{I}_D; \alpha_d \ne 0 \quad \forall d \in \{1, \ldots, D\}\}$$

is empty or finite. If this set is non-empty, any element α in this set satisfies

$$\forall d \in \{1, \ldots, D\} \quad \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X, X')}{\sum_{j=1}^D \alpha_j q_j(X, X')} \right) = 1,$$

which implies

$$0 = \sum_{d=1}^{D} \alpha_d^{max} \left(\mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X,X')}{\sum_{j=1}^{D} \alpha_j q_j(X,X')} \right) - 1 \right)$$
$$= \mathbb{E}_{\bar{\pi}} \left(\frac{\sum_{d=1}^{D} \alpha_d^{max} q_d(X,X')}{\sum_{j=1}^{D} \alpha_j q_j(X,X')} - 1 \right) \geq \mathbb{E}_{\bar{\pi}} \left(\log \frac{\sum_{d=1}^{D} \alpha_d^{max} q_d(X,X')}{\sum_{j=1}^{D} \alpha_j q_j(X,X')} \right) \geq 0$$

By unicity of the global maximum of $\mathcal{E}_{\bar{\pi}}$, we conclude that $\alpha = \alpha^{max}$ and hence **iii)**. □

Proposition 4.1 implies that our recursive procedure satisfies $\mathcal{E}_{\bar{\pi}}(\alpha^{t+1}) \geq \mathcal{E}_{\bar{\pi}}(\alpha^t)$. Therefore, the Kullback Leibler divergence criterion (2) decreases at each step. This property is closely linked with the EM algorithm (Robert and Casella, 2004, Section 5.3). More precisely, consider the mixture model

$$V \sim \mathcal{M}(1, (\alpha_1, \ldots, \alpha_D)) \quad \text{and} \quad W = (X, X')|V \sim \pi(dx) Q_V(x, dx')$$

with parameter α. We denote by $\bar{\mathbb{E}}_\alpha$ the corresponding expectation, by $p_\alpha(v,w)$ the joint density of (V,W) wrt $\mu \otimes \mu$ and by $p_\alpha(w)$ the density of W wrt μ. Then it is easy to check that $\mathcal{E}_{\bar{\pi}}(\alpha) = \int \log(p_\alpha(w)) \bar{\pi}(dw)$ which is an average version of the criterion to be maximized in the EM algorithm when only W is observed. In that case, a natural idea adapted from the EM algorithm would be to update α according to the iterative scheme

$$\alpha^{t+1} = \arg\max_{\alpha \in \mathscr{S}} \int \bar{\mathbb{E}}_{\alpha^t} \left[\log p_\alpha(V,w) | \, w \right] \bar{\pi}(dw) \, .$$

By direct algebra, this definition of α^{t+1} is equivalent to the update formula $\alpha^{t+1} = F(\alpha^t)$ that we used above. Our algorithm then appears as an averaged EM, but preserves the deterministic increase of the criterion enjoyed by EM.

The following proposition ensures that any α different from α^{max} is repulsive.

Proposition 4.2. *Under **(A1)** and **(A2)**, for every $\alpha \in \mathscr{S} \setminus \{\alpha^{max}\}$, there exists a neighborhood V_α of α such that, if $\alpha^{t_0} \in V_\alpha$, then $(\alpha^t)_{t \geq t_0}$ leaves V_α within a finite time.*

Proof. Let $\alpha \in \mathscr{S} \setminus \{\alpha^{max}\}$. Then, using the inequality $u - 1 \geq \log u$,

$$\sum_{d=1}^{D} \alpha_d^{max} \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X,X')}{\sum_{j=1}^{D} \alpha_j q_j(X,X')} \right) - 1 \geq \mathbb{E}_{\bar{\pi}} \left(\log \frac{\sum_{d=1}^{D} \alpha_d^{max} q_d(X,X')}{\sum_{j=1}^{D} \alpha_j q_j(X,X')} \right) > 0$$

which implies that there exists $d \in \{1, \ldots, D\}$ such that

$$\mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X,X')}{\sum_{j=1}^{D} \alpha_j q_j(X,X')} \right) > 1.$$

Let $(W_n)_{n \geq 0}$ be a non increasing sequence of neighborhoods of α in $\mathscr{S}$. We have by the monotone convergence theorem,

$$\begin{aligned} 1 < \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X,X')}{\sum_{j=1}^{D} \alpha_j q_j(X,X')} \right) &= \mathbb{E}_{\bar{\pi}} \left(\lim_{n \to \infty} \inf_{\beta \in W_n} \frac{q_d(X,X')}{\sum_{j=1}^{D} \beta_j q_j(X,X')} \right) \\ &= \lim_{n \to \infty} \mathbb{E}_{\bar{\pi}} \left(\inf_{\beta \in W_n} \frac{q_d(X,X')}{\sum_{j=1}^{D} \beta_j q_j(X,X')} \right) \\ &\leq \lim_{n \to \infty} \inf_{\beta \in W_n} \mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X,X')}{\sum_{j=1}^{D} \beta_j q_j(X,X')} \right) \end{aligned}$$

Thus, there exist $W_{n_0} = V_\alpha$ a neighborhood of α and $\eta > 1$ such that for all $\beta \in V_\alpha$,

$$\mathbb{E}_{\bar{\pi}} \left(\frac{q_d(X,X')}{\sum_{j=1}^{D} \beta_j q_j(X,X')} \right) > \eta. \tag{6}$$

Now use that for all $t \geq 0$ and $d \in \{1, \ldots, D\}$, $1 \geq \alpha_d^t > 0$ and combine (6) with the update formulas for α_d^t (given by (3)). This shows that $(\alpha^t)_{t \geq 0}$ leaves V_α within a finite time. □

We thus conclude that the maximization algorithm is convergent:

Proposition 4.3. *Under (A1) and (A2),*

$$\lim_{t\to\infty} \alpha^t = \alpha^{max}\,.$$

Proof. First, note that $\mathcal{I}_D$ is a finite set which contains α^{max}. Write $\mathcal{I}_D = \{\beta_0, \beta_1, \ldots, \beta_I\}$ with $\beta_0 = \alpha^{max}$. If we introduce a sequence $(W_i)_{0\le i\le I}$ of disjoint neighborhoods of the β_i's so that for all $0 \le i \le I$, $F(W_i)$ is disjoint from $\cup_{j\neq i} W_j$ (this is possible since $F(\beta_i) = \beta_i$ and F is continuous) and, for all $i \in \{1, \ldots, I\}$, $W_i \subset V_{\beta_i}$ where the (V_{β_i})'s are defined in the proof of Proposition 4.2.

The sequence $(\mathcal{E}_{\bar\pi}(\alpha^t))_{t\ge 0}$ is upper-bounded and non decreasing and therefore it converges. This implies that $\lim_{t\to\infty} \mathcal{E}_{\bar\pi} \circ F(\alpha^t) - \mathcal{E}_{\bar\pi}(\alpha^t) = 0$. By continuity of $\mathcal{E}_{\bar\pi} \circ F - \mathcal{E}_{\bar\pi}$, there exists $T > 0$ such that for all $t \ge T$, $\alpha_t \in \cup_j W_j$. Since $F(W_i)$ is disjoint from $\cup_{j\neq i} W_j$, this implies that there exists $i \in \{0, \ldots, I\}$ such that for all $t \ge T$, $\alpha^t \in W_i$. By Proposition 4.2, i cannot be in $\{1, \ldots, I\}$. Thus, for all $t \ge T$, $\alpha^t \in W_0$ which is a neighborhood of $\beta_0 = \alpha^{max}$. The proof is completed. □

5 The Rao-Blackwellized D-kernel PMC

The update formula (3) has been shown to improve the Kullback Leibler divergence criterion at every iteration. We now discuss how to implement this mechanism within a PMC algorithm that resembles the previous D-kernel algorithm. The only difference with the algorithm of Section 3.1 is that we make use of the kernel structure in the computation of the importance weight: in MCMC terminology, this is called "Rao-Blackwellization" (Robert and Casella, 2004, Section 4.2) and it is known to provide variance reduction in data augmentation settings (Robert and Casella, 2004, Section 9.2). In the current context, the improvement brought by Rao-Blackwellization is dramatic, in that the modified algorithm does converge to the proposal mixture that is closest to the target distribution in the sense of the Kullback Leibler divergence. More precisely, a Monte Carlo version of the update formula (3) can be implemented in the iterative definition of the mixture weights, in the same way as MCEM approximates EM (Robert and Casella, 2004, Section 5.3.3).

5.1 The algorithm

In importance sampling as well as in MCMC settings, the conditioning improvement brought by Rao-Blackwellization may be significant (Celeux et al., 2003). In the context of the D-kernel PMC scheme, the Rao-Blackwellization argument is that it is not necessary to use the mixture component in the computation of the importance weight but rather the whole mixture. The importance weight is therefore

$$\pi(x_{i,t}) \Big/ \sum_{d=1}^{D} \alpha_d^{t,N} q_d(\tilde x_{i,t-1}, x_{i,t}) \quad \text{rather than} \quad \pi(x_{i,t}) / q_{K_{i,t}}(\tilde x_{i,t-1}, x_{i,t})$$

as in the algorithm of Section 3.1. As already noted by Hesterberg (1998), the use of the whole mixture in the importance weight provides a robust tool for preventing infinite variance importance sampling estimators. In our setup, this choice of weight will guarantee that the following algorithm converges to the optimal mixture.

Rao-Blackwellized D-kernel PMC algorithm:
At time 0, use the same step as in the generic PMC algorithm to produce the sample $(\tilde x_{i,0}, J_{i,0})_{1\le i\le N}$ and set $\alpha_d^{1,N} = 1/D$ for all $1 \le d \le D$.
At time t $(t = 1, \ldots, T)$,

a) Conditionally on $\sigma\left\{(\alpha_d^{t,N})_{1\le d\le D}\right\}$, generate

$$(K_{i,t})_{1\le i\le N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\alpha_d^{t,N})_{1\le d\le D})$$

b) Conditionally on $\sigma\{(\tilde x_{i,t-1}, K_{i,t})_{1\le i\le N}\}$, generate independent

$$(x_{i,t})_{1\le i\le N} \sim Q_{K_{i,t}}(\tilde x_{i,t-1}, \cdot)$$

and set $\omega_{i,t} = \pi(x_{i,t}) \Big/ \sum_{d=1}^{D} \alpha_d^{t,N} q_d(\tilde x_{i,t-1}, x_{i,t})$;

c) Conditionally on $\sigma\{(\tilde{x}_{i,t-1}, K_{i,t}, x_{i,t})_{1\le i\le N}\}$, generate

$$(J_{i,t})_{1\le i\le N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\bar{\omega}_{i,t})_{1\le i\le N})$$

and set $(1 \le i \le N,\ 1 \le d \le D)$

$$\tilde{x}_{i,t} = x_{J_{i,t},t}\,, \quad \alpha_d^{t+1,N} = \sum_{i=1}^{N} \bar{\omega}_{i,t} \mathbb{I}_d(K_{i,t})$$

Note that, once more, the adaptive mechanism is based on the importance weights. The update of the α_d's in Step c) is the Monte Carlo version of (3) and we now show that this algorithm is converging.

5.2 The LLN for the Rao-Blackwellized D-kernel PMC algorithm

Not very surprisingly, the population of points obtained at each iteration of the Rao-Blackwellized algorithm above approximates the target distribution in the sense of the weak Law of Large Numbers (LLN). Note that the convergence holds under the very weak assumption (**A1**) and for any test function h that is absolutely integrable wrt the target distribution π. The function h may thus be unbounded.

Theorem 5.1. *Under (**A1**), for any function h in L_π^1 and for all $t \ge 0$,*

$$\sum_{i=1}^{N} \bar{\omega}_{i,t} h(x_{i,t}) \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} \pi(h) \tag{7}$$

$$\frac{1}{N}\sum_{i=1}^{N} h(\tilde{x}_{i,t}) \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} \pi(h) \tag{8}$$

Proof. We proceed by induction wrt t on the two limiting results (8) and (7). The case $t = 0$ is the basic importance sampling convergence result. Now, let $t \ge 1$ and assume that (8) and (7) both hold for $t-1$. We will just show (7) since (8) is a straightforward consequence of (7) and Theorem A.1 due to multinomial sampling, by noting that

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} h(\tilde{x}_{i,t}) \middle| (x_{i,t})_{1\le i\le N} \right) = \sum_{i=1}^{N} \omega_{i,t} h(x_{i,t}).$$

To prove (7), we will check that

$$\frac{1}{N}\sum_{i=1}^{N} \omega_{i,t} h(x_{i,t}) \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} \pi(h)$$

$$\frac{1}{N}\sum_{i=1}^{N} \omega_{i,t} \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} 1.$$

The later limit is also a direct consequence of the former with $h = 1$. We apply Theorem A.1 with $\mathcal{G}_N = \sigma\left((\tilde{x}_{i,t-1})_{1\le i\le N}, (\alpha_d^{t,N})_{1\le d\le D}\right)$ and $U_{N,i} = N^{-1}\omega_{i,t}h(x_{i,t})$. Conditionally on $\mathcal{G}_N$, the $(x_{i,t})_{1\le i\le N}$'s are independent and

$$x_{i,t}|\mathcal{G}_N \sim \sum_{d=1}^{D} \alpha_d^{t,N} Q_d(\tilde{x}_{i,t-1}, \cdot)$$

Noting that

$$\sum_{i=1}^{N} \mathbb{E}\left(\frac{\omega_{i,t}h(x_{i,t})}{N} \middle| \mathcal{G}_N \right) = \sum_{i=1}^{N} \mathbb{E}\left(\frac{\pi(x_{i,t})h(x_{i,t})}{N\sum_{d=1}^{D} \alpha_d^{t,N} q_d(\tilde{x}_{i,t-1}, x_{i,t})} \middle| \mathcal{G}_N \right) = \pi(h),$$

to apply Theorem A.1, we only need to check condition (iii). Now, write

$$\begin{aligned}
&\sum_{i=1}^{N} \mathbb{E}\left(\frac{\omega_{i,t} h(x_{i,t})}{N} \mathbb{I}_{\{\omega_{i,t} h(x_{i,t}) > C\}} \middle| \mathcal{G}_N \right) \\
&\quad = \frac{1}{N} \sum_{i=1}^{N} \int \pi(dx) h(x) \mathbb{I}_{\left\{ \frac{\pi(x)h(x)}{\sum_{d=1}^{D} \alpha_d^{t,N} q_d(\tilde{x}_{i,t-1}, x)} > C \right\}} \\
&\quad \leq \sum_{d=1}^{D} \frac{1}{N} \sum_{i=1}^{N} \int \pi(dx) h(x) \mathbb{I}_{\left\{ \frac{\pi(x)h(x)}{D^{-1} q_d(\tilde{x}_{i,t-1}, x)} > C \right\}}
\end{aligned}$$

Note $F_C(u) = \int \pi(dx) h(x) \mathbb{I}_{\left\{ \frac{\pi(x)h(x)}{D^{-1} q_d(u,x)} > C \right\}}$. We have $F_C(u) \leq \pi(h)$ and thus, by the induction assumption

$$N^{-1} \sum_{i=1}^{N} F_C(\tilde{x}_{i,t-1}) \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} \pi(F_C).$$

The proof is completed since

$$\pi(F_C) = \iint \pi(dx)\pi(dx') h(x) \mathbb{I}_{\left\{ \frac{\pi(x)h(x)}{D^{-1} q_d(x',x)} > C \right\}} \overset{C\to\infty}{\longrightarrow} 0 .$$

□

5.3 Convergence of the weights

The next proposition ensures that, at each iteration of the algorithm, the population of points is modified according to a mixture of kernels whose weights approximate the ones obtained by the iterative procedure described in Section 4 for minimizing the Kullback divergence criterion.

Proposition 5.1. *Under* ***(A1)****, for all* $t \geq 1$,

$$\forall 1 \leq d \leq D, \quad \alpha_d^{t,N} \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} \alpha_d^t \tag{9}$$

where the α_d^t*'s are defined in (3).*

Combining Proposition 5.1 with Proposition 4.3, we obtain that, under assumptions **(A1)** and **(A2)**, the Rao–Blackwellized version of the PMC algorithm automatically adapts the weights of the proposed mixture of kernels and converges to the optimal combination of mixtures wrt to the Kullback divergence criterion defined in Section 4.

Proof. The case $t = 1$ is obvious. Now, assume (9) holds for some $t \geq 1$. As in the proof of Proposition 3.1, we now prove that

$$\begin{aligned}
&\frac{1}{N} \sum_{i=1}^{N} \omega_{i,t} \mathbb{I}_d(K_{i,t}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(x_{i,t})}{\sum_{l=1}^{D} \alpha_l^{t,N} q_l(\tilde{x}_{i,t-1}, x_{i,t})} \mathbb{I}_d(K_{i,t}) \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} \alpha_d^{t+1}, \\
&\frac{1}{N} \sum_{i=1}^{N} \omega_{i,t} \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} 1.
\end{aligned}$$

Only the first convergence needs be considered since the latter can be easily deduced from the former.
We apply Theorem A.1 with $\mathcal{G}_N = \sigma\left((\tilde{x}_{i,t-1})_{1\leq i\leq N}, (\alpha_d^{t,N})_{1\leq d\leq D} \right)$ and $U_{N,i} = N^{-1} \omega_{i,t} \mathbb{I}_d(K_{i,t})$. Conditionally on $\mathcal{G}_N$, $(K_{i,t}, x_{i,t})_{1\leq i\leq N}$ are independent and for all (d, A) in $\{1, \ldots, D\} \times \mathcal{A}$,

$$\mathbb{P}\left(K_{i,t} = d, x_{i,t} \in A \middle| \mathcal{G}_N \right) = \alpha_d^{t,N} Q_d(\tilde{x}_{i,t-1}, A)$$

To apply Theorem A.1, we just need to check condition (iii). We have

$$\begin{aligned}
&\mathbb{E}\left(\sum_{i=1}^{N}\frac{\omega_{i,t}\mathbb{I}_d(K_{i,t})}{N}\mathbb{I}_{\{\omega_{i,t}\mathbb{I}_d(K_{i,t})>C\}}\middle|\mathcal{G}_N\right)\\
&\leq\sum_{j=1}^{D}\frac{1}{N}\sum_{i=1}^{N}\int\pi(dx)\frac{\alpha_d^{t,N}q_d(\tilde{x}_{i,t-1},x)}{\sum_{l=1}^{D}\alpha_l^{t,N}q_l(\tilde{x}_{i,t-1},x)}\mathbb{I}_{\left\{\frac{\pi(x)}{D^{-1}q_j(\tilde{x}_{i,t-1},x)}>C\right\}}\\
&\leq\sum_{j=1}^{D}\frac{1}{N}\sum_{i=1}^{N}\int\pi(dx)\mathbb{I}_{\left\{\frac{\pi(x)}{D^{-1}q_j(\tilde{x}_{i,t-1},x)}>C\right\}}\overset{N\to\infty}{\longrightarrow_{\mathbb{P}}}\sum_{j=1}^{D}\bar{\pi}\left(\frac{\pi(x)}{D^{-1}q_j(x',x)}>C\right)
\end{aligned}$$

by the LLN stated in Theorem 5.1. The rhs converges to 0 as C grows to infinity since by assumption **(A1)**, $\bar{\pi}\{q_j(x,x')=0\}=0$. Thus, Theorem A.1 applies and

$$\frac{1}{N}\sum_{i=1}^{N}\omega_{i,t}\mathbb{I}_d(K_{i,t})-\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\omega_{i,t}\mathbb{I}_d(K_{i,t})\middle|\mathcal{G}_N\right)\overset{N\to\infty}{\longrightarrow_{\mathbb{P}}}0.$$

To complete the proof, it remains to show that

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\omega_{i,t}\mathbb{I}_d(K_{i,t})\middle|\mathcal{G}_N\right)=\frac{1}{N}\sum_{i=1}^{N}\int\pi(dx)\frac{\alpha_d^{t,N}q_d(\tilde{x}_{i,t-1},x)}{\sum_{l=1}^{D}\alpha_l^{t,N}q_l(\tilde{x}_{i,t-1},x)}\overset{N\to\infty}{\longrightarrow_{\mathbb{P}}}\alpha_d^{t+1}\tag{10}$$

Using again the LLN stated in Theorem 5.1,

$$\frac{1}{N}\sum_{i=1}^{N}\int\pi(dx)\frac{\alpha_d^{t}q_d(\tilde{x}_{i,t-1},x)}{\sum_{l=1}^{D}\alpha_l^{t}q_l(\tilde{x}_{i,t-1},x)}\overset{N\to\infty}{\longrightarrow_{\mathbb{P}}}\mathbb{E}_{\bar{\pi}}\left(\frac{\alpha_d^{t}q_d(X,X')}{\sum_{l=1}^{D}\alpha_l^{t}q_l(X,X')}\right)=\alpha_d^{t+1}\tag{11}$$

Thus, to prove (10), we use (11) and check that the difference between both terms converges to 0 in probability. To see this, first note that for all $t\geq 1$, for all d in $\{1,\ldots,D\}$, $\alpha_d^t>0$ and thus, by the induction assumption, for all d in $\{1,\ldots,D\}$, $\frac{\alpha_d^{t,N}-\alpha_d^t}{\alpha_d^t}\overset{N\to\infty}{\longrightarrow_{\mathbb{P}}}0$. Using that $\left|\frac{A}{B}-\frac{C}{D}\right|\leq\left|\frac{A}{B}\right|\left|\frac{D-B}{D}\right|+\left|\frac{A-C}{C}\right|\left|\frac{C}{D}\right|$, we have by straightforward algebra,

$$\begin{aligned}
&\left|\frac{\alpha_d^{t,N}q_d(\tilde{x}_{i,t-1},x)}{\sum_{l=1}^{D}\alpha_l^{t,N}q_l(\tilde{x}_{i,t-1},x)}-\frac{\alpha_d^{t}q_d(\tilde{x}_{i,t-1},x)}{\sum_{l=1}^{D}\alpha_l^{t}q_l(\tilde{x}_{i,t-1},x)}\right|\\
&\leq\frac{\alpha_d^{t,N}q_d(\tilde{x}_{i,t-1},x)}{\sum_{j=1}^{D}\alpha_l^{t,N}q_j(\tilde{x}_{i,t-1},x)}\left(\sup_{l\in\{1,\ldots,D\}}\left|\frac{\alpha_l^{t,N}-\alpha_l^{t}}{\alpha_l^{t}}\right|\right)\\
&\qquad+\left|\frac{\alpha_d^{t,N}-\alpha_d^{t}}{\alpha_d^{t}}\right|\frac{\alpha_d^{t}q_d(\tilde{x}_{i,t-1},x)}{\sum_{l=1}^{D}\alpha_l^{t}q_l(\tilde{x}_{i,t-1},x)}\\
&\leq 2\sup_{l\in\{1,\ldots,D\}}\left|\frac{\alpha_l^{t,N}-\alpha_l^{t}}{\alpha_l^{t}}\right|.
\end{aligned}$$

The proof follows from $\frac{\alpha_d^{t,N}-\alpha_d^t}{\alpha_d^t}\overset{N\to\infty}{\longrightarrow_{\mathbb{P}}}0$. □

5.4 The CLT for the Rao-Blackwellized D-kernel PMC algorithm

We now state and prove a CLT for the weighed and the unweighted samples when the size of the population grows to infinity. As noted in the SIR algoritm (see Section 2.2), the asymptotic variance associated with the unweighted sample is larger than the variance of the weighted sample.

Theorem 5.2. *Under **(A1)**,*

i) *For all h satisfying $\bar{\pi}\left(h^2(x')\frac{\pi(x)}{q_d(x,x')}\right)<\infty$ for some $d\in\{1,\ldots,D\}$, we have*

$$\sqrt{N}\sum_{i=1}^{N}\bar{\omega}_{i,t}\left\{h(x_{i,t})-\pi(h)\right\}\overset{\mathcal{L}}{\longrightarrow}\mathcal{N}(0,\sigma_t^2),\tag{12}$$

with $\sigma_t^2=\bar{\pi}\left((h(x')-\pi(h))^2\frac{\pi(x')}{\sum_{d=1}^{D}\alpha_d^t q_d(x,x')}\right)$.

ii) *If moreover* $\pi(h^2) < \infty$, *then*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (h(\tilde{x}_{i,t}) - \pi(h)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_t^2 + \mathbb{V}_\pi(h)) \tag{13}$$

Note that amongst the conditions under which this theorem applies, the integrability condition

$$\bar{\pi}\left(h^2(x') \frac{\pi(x)}{q_d(x,x')}\right) < \infty \tag{14}$$

is required for *some* d in $\{1, \ldots, D\}$ and not for *all* d. Thus, situations where some transition kernels $q_d(\cdot,\cdot)$ do not satisfy (14) can still be covered by this theorem provided that (14) holds for at least one particular kernel. An equivalent expression of the asymptotic variance σ_t^2 is

$$\sigma_t^2 = \mathbb{V}_\nu\left((h - \pi(h))\frac{\bar{\pi}}{\nu}\right) \text{ where } \nu(dx, dx') = \pi(dx)\left(\sum_{d=1}^{D} \alpha_d^t Q_d(x, dx')\right).$$

Written as above, σ_t^2 turns to have the same expression as the asymptotic variance that appears in the CLT associated to the self-normalized IS algorithm (SNIS) (see Section 2.1 for a description of the algorithm and the associated CLT) where the proposal distribution is ν and the target distribution $\bar{\pi}$. Still, the SNIS algorithm can not be implemented here since by the above definition of ν, the proposal distribution depends on both π and the weights (α_d^t) which are unknown.

Proof. Without loss of generality, we may assume that $\pi(h) = 0$.
Let $d_0 \in \{1, \ldots, D\}$ such that $\bar{\pi}\left(h^2(x') \frac{\pi(x)}{q_{d_0}(x,x')}\right) < \infty$. In the proof of Theorem 5.1, it has been shown that $\frac{1}{N}\sum_{i=1}^{N} \omega_{i,t} \xrightarrow{N\to\infty}_{\mathbb{P}} 1$ and thus, we only need to prove that

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \omega_{i,t} h(x_{i,t}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_t^2) \tag{15}$$

We will apply Theorem A.2 with

$$U_{N,i} = \frac{1}{\sqrt{N}} \omega_{i,t} h(x_{i,t}) = \frac{1}{\sqrt{N}} \frac{\pi(x_{i,t}) h(x_{i,t})}{\sum_{d=1}^{D} \alpha_d^{t,N} q_d(\tilde{x}_{i,t-1}, x_{i,t})},$$
$$\mathcal{G}_N = \sigma\left\{(\tilde{x}_{i,t-1})_{1\le i\le N}, (\alpha_d^{t,N})_{1\le d\le D})\right\}.$$

Conditionally on $\mathcal{G}_N$, the $(x_{i,t})_{1\le i\le N}$ are independent and

$$x_{i,t} | \mathcal{G}_N \sim \sum_{d=1}^{D} \alpha_d^{t,N} Q_d(\tilde{x}_{i,t-1}, \cdot).$$

Conditions (i) and (ii) of Theorem A.2 are straightforwardly satisfied. To check condition (iii), first note that $\mathbb{E}(U_{N,i} | \mathcal{G}_N) = \pi(h) = 0$. Moreover,

$$A_N = \sum_{i=1}^{N} \mathbb{E}(U_{N,i}^2 | \mathcal{G}_N) = \frac{1}{N} \sum_{i=1}^{N} \int \pi(dx) h^2(x) \frac{\pi(x)}{\sum_{d=1}^{D} \alpha_d^{t,N} q_d(\tilde{x}_{i,t-1}, x)}$$

By the LLN for $(\tilde{x}_{i,t})$ stated in Theorem 5.1, we have

$$B_N = \frac{1}{N} \sum_{i=1}^{N} \int \pi(dx) h^2(x) \frac{\pi(x)}{\sum_{d=1}^{D} \alpha_d^{t} q_d(\tilde{x}_{i,t-1}, x)} \xrightarrow{N\to\infty}_{\mathbb{P}} \sigma_t^2$$

To prove that condition (iii) holds, it is thus sufficient to show that $|B_N - A_N| \xrightarrow{N\to\infty}_{\mathbb{P}} 0$. Since $\alpha_{d_0}^{t,N} \xrightarrow{N\to\infty}_{\mathbb{P}} \alpha_{d_0}^t > 0$, it is sufficient to consider the bound

$$\begin{aligned}
&\mathbb{I}_{\{\alpha_{d_0}^{t,N} > 2^{-1}\alpha_{d_0}^t\}} |B_N - A_N| \\
&\quad \leq \mathbb{I}_{\{\alpha_{d_0}^{t,N} > 2^{-1}\alpha_{d_0}^t\}} \sup_{1\leq d\leq D} \left(\frac{\alpha_d^t - \alpha_d^{t,N}}{\alpha_d^t}\right) \frac{1}{N}\sum_{j=1}^N \int \pi(dx) \frac{h^2(x)\pi(x)}{\sum_{d=1}^D \alpha_d^{t,N} q_d(\tilde{x}_{i,t-1}, x)} \\
&\quad \leq \left(\sup_{1\leq d\leq D} \left(\frac{\alpha_d^t - \alpha_d^{t,N}}{\alpha_d^t}\right) \frac{1}{N}\sum_{j=1}^N \int \pi(dx) \frac{h^2(x)\pi(x)}{2^{-1}\alpha_{d_0}^t q_{d_0}(\tilde{x}_{i,t-1}, x)} \right) \xrightarrow{N\to\infty}_{\mathbb{P}} 0
\end{aligned}$$

Thus, condition (iii) is satisfied. Now, consider condition (iv). Using the same argument as in condition (iii), we consider

$$\begin{aligned}
&\mathbb{I}_{\{\alpha_{d_0}^{t,N} > 2^{-1}\alpha_{d_0}^t\}} \sum_{i=1}^N \mathbb{E}\left(\frac{1}{N}\omega_{i,t}^2 h^2(x_{i,t}) \mathbb{I}_{\left\{\frac{\pi(x_{i,t})h(x_{i,t})}{\sum_{d=1}^D \alpha_d^{t,N} q_d(\tilde{x}_{i,t-1}, x_{i,t})} > C\right\}} \,\middle|\, \mathcal{G}_N \right) \\
&\quad \leq \frac{1}{N}\sum_{i=1}^N \int \pi(dx) h^2(x) \frac{\pi(x)}{2^{-1}\alpha_{d_0}^t q_{d_0}(\tilde{x}_{i,t-1}, x)} \mathbb{I}_{\left\{\frac{\pi(x)h(x)}{2^{-1}\alpha_{d_0}^t q_{d_0}(\tilde{x}_{i,t-1}, x)} > C\right\}} \\
&\quad \xrightarrow{N\to\infty}_{\mathbb{P}} \bar{\pi}\left(h^2(x) \frac{\pi(x)}{2^{-1}\alpha_{d_0}^t q_{d_0}(x', x)} \mathbb{I}_{\left\{\frac{\pi(x)h(x)}{2^{-1}\alpha_{d_0}^t q_{d_0}(x', x)} > C\right\}} \right)
\end{aligned}$$

which converges to 0 as C grows to infinity. Thus, Theorem A.2 applies and the proof of (12) is completed. The proof of (13) is derived as a direct application of Theorem A.2 as in the SIR result by setting $U_{N,i} = \frac{1}{\sqrt{N}} h(\tilde{x}_{i,t})$ and $\mathcal{G}_N = \sigma\left((x_{i,t})_{1\leq i\leq N}, (\omega_{i,t})_{1\leq i\leq N}\right)$. □

6 Illustrations

In this section, we briefly show how the iterations of the PMC algorithm quickly implement adaptivity towards the most efficient mixture of kernels though three examples of moderate difficulty. (The R programs are available on the authors' websites.)

Example 1. As a first toy example, consider the case of the target π being the density of a normal mixture

$$\sum_{i=1}^3 \frac{1}{3}\,\mathcal{N}(\mu_i, \sigma_i^2) \tag{16}$$

and of a independent normal mixture proposal with the same means and variances as in (16) but started with different weights $\alpha_d^{0,N}$. Note that this is a very special case of D kernel PMC scheme in that the Markov kernels of Section 5.1 are then independent proposals. In this case, the optimal choice of weights is obviously $\alpha_d^\star = 1/3$. In our experiment, we used $\mu_1 = -2$, $\mu_2 = 0$, $\mu_3 = 2$ and $\sigma_1 = 1/3$, $\sigma_2 = 2/3$, $\sigma_3 = 1$. The starting values $\alpha_d^{0,N}$ are indicated on the left of Figure 1, which clearly shows the convergence to the optimal values 1/3 and 2/3 for the two first cumulated weights in less than 10 iterations. (Generating more simulated points at each iteration do stabilize the convergence graph but the primary aim of this example is to exhibit the fast convergence to the true optimal values of the weights.)

Example 2. As a second toy example, consider the case of a $\mathcal{N}(0,1)$ target and of the following mixture of D kernels

$$\alpha_1^{t,N}\mathscr{T}_2(\tilde{x}_{i,t-1}, 1) + \alpha_2^{t,N}\mathcal{N}(\tilde{x}_{i,t-1}, \sigma_2^2) + \alpha_3^{t,N}\mathcal{N}(\tilde{x}_{i,t-1}, \sigma_3^2)\,, \tag{17}$$

where $\sigma_2^2 = 4$ and $\sigma_3^2 = 1/4$. (The first proposal in the mixture is thus a Student $\mathscr{T}_2$ distribution centered at the current value $\tilde{x}_{i,t-1}$.) Figure 2 details the convergence of the weights to the optimal values for several starting values. (Note that the optimal values can be approximated numerically by a discretization of the simplex in $\mathbb{R}^3$. For the discretization step adopted in Figure 2, the optimum corresponds to $\alpha_1^\star = 0.41$ and $\alpha_2^\star = 0.51$.) While the sequences of weights $(\alpha_1^{t,N}, \alpha_2^{t,N})$ do not always converge exactly to the same value, this is due to the considerable flatness of the Kullback–Leibler divergence in this region.

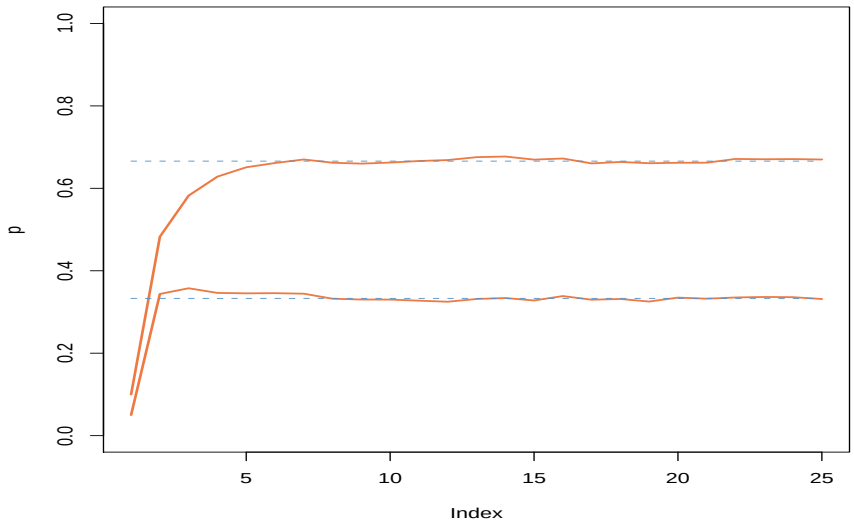

Figure 1: Convergence of the cumulated weights $\alpha_1^{t,N}$ and $\alpha_1^{t,N} + \alpha_2^{t,N}$ for the three component normal mixture to the optimal values 1/3 and 2/3 (represented by dotted lines). At each iteration, $N = 10,000$ points were simulated from the D-kernel proposal.

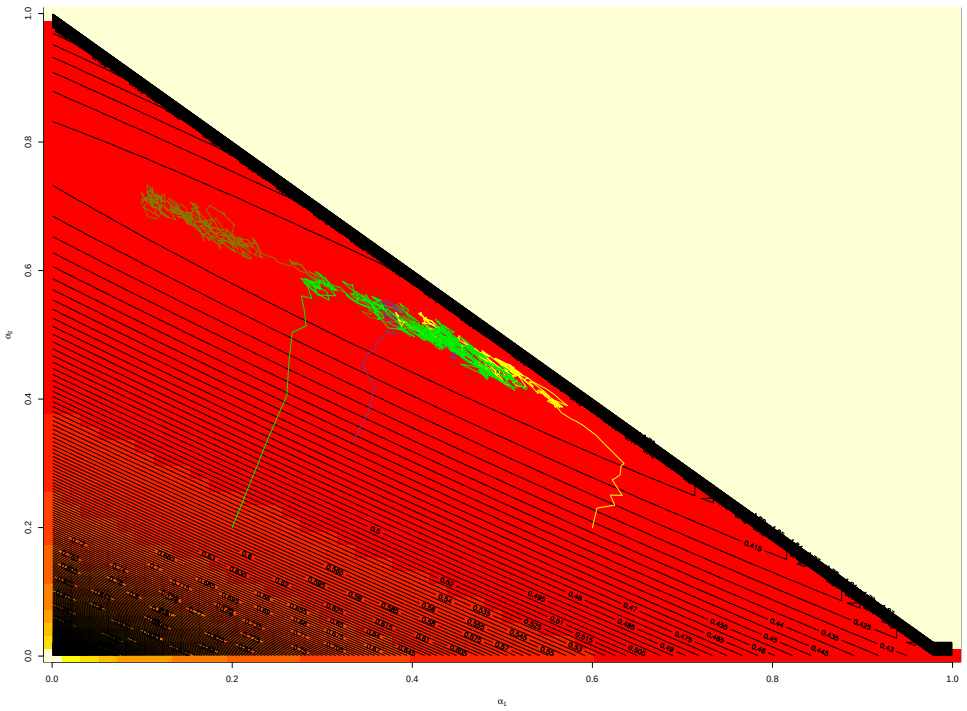

Figure 2: Numerical approximation of the Kullback–Leibler divergence for the three component mixture proposal (17) in the simplex of $\mathbb{R}^3$. (The discretization step is 1/75 in both α_1 and α_2 directions.) Superposition of the path of four calls to the D-kernel PMC algorithm when started from different values $(\alpha_1^{1,N}, \alpha_2^{1,N})$. The number T of iterations is between 150 and 500 depending on the starting values, while the sample size is 50,000.

	0	1	Total
0	60	364	424
1	36	240	276
Total	96	604	700

Table 1: Two-by-two contingency table.

Example 3. Our third example is a contingency table inspired from Agresti (2002), given in Table 1. We model this dataset by a Poisson regression,

$$x_{ij} \sim \mathscr{P}\left(\exp\left(\alpha_i + \beta_j\right)\right) \qquad (i, j = 0, 1)\,,$$

with $\alpha_0 = 0$ for identifiability reasons. We use a flat prior on the parameter $\theta = (\alpha_1, \beta_0, \beta_1)$ and run the PMC D-kernel algorithm with a mixture of 10 normal random walk proposals, $\mathscr{N}(\tilde{\theta}_{i,t-1}, \varrho_d I(\hat{\theta}))$ $(d = 1, \ldots, 10)$, where $I(\hat{\theta}))$ is the Fisher information matrix evaluated at the MLE, $\hat{\theta} = (-0.43, 4.06, 5.9)$ and where the scales ϱ_d vary from $1.35e-19$ to $1.54e+07$ (the ϱ_d's are equidistributed on a logarithmic scale). The result of 5 (successive) iterations of the Rao-Blackwell D-kernel algorithm is as follows: unsurprisingly, the largest variance kernels are hardly ever sampled but fulfill their main role of variance stabilizers in the importance sampling weights while the mixture concentrates on the medium variances, with a quick convergence of the mixture weights to the limiting weights. This convergence is illustrated in Figure 4 for the cumulated weights of the 5th, 6th, 7th and 8th components of the mixture, which converge to 0, 0.003, 0.259 and 0.738, respectively. The adequation of the simulated sample with the target distribution is shown in Figure 3, since the points of the sample do coincide with the (unique) modal region of the posterior distribution. The last row of Figure 3 (see also the log-posterior histograms in Figure 5) shows in addition that there is no degeneracy in the produced samples: most points in the last sample have very similar posterior values. For instance, 20% of the sample corresponds to 95% of the weights, while 1% of the sample corresponds to 31% of the weights. A closer look at convergence is provided by Figure 5 where the histograms of the resampled samples are represented, along with the distribution of the loglikelihood and the empirical cdf of the importance weights: they do not signal any degeneracy phenomenon but on the opposite a clear stabilization around the values of interest.

Acknowledgments

The authors are grateful to Olivier Cappé, Paul Fearnhead and Eric Moulines for helpful comments and discussions. This work was partially supported by an ACI *"Nouvelles Interfaces des Mathématiques"* grant from the Ministère de la Recherche.

References

Agresti, A. (2002). *Categorical Data Analysis.* John Wiley, second edition.

Andrieu, C. and Moulines, E. (2004). On the ergodicity properties of some adaptive MCMC algorithms. Technical report, Department of Mathematics and Statistics, University of Bristol, UK.

Andrieu, C. and Robert, C. (2001). Controlled Markov chain Monte Carlo methods for optimal sampling. Technical Report 0125, Univ. Paris Dauphine.

Cappé, O., Guillin, A., Marin, J., and Robert, C. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929.

Cappé, O., Moulines, E., and Rydèn, T. (2005). *Inference in Hidden Markov Models.* Springer-Verlag, New York.

Celeux, G., Marin, J., and Robert, C. (2003). Iterated importance sampling in missing data problems. Technical report, Université Paris Dauphine.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.* (to appear).

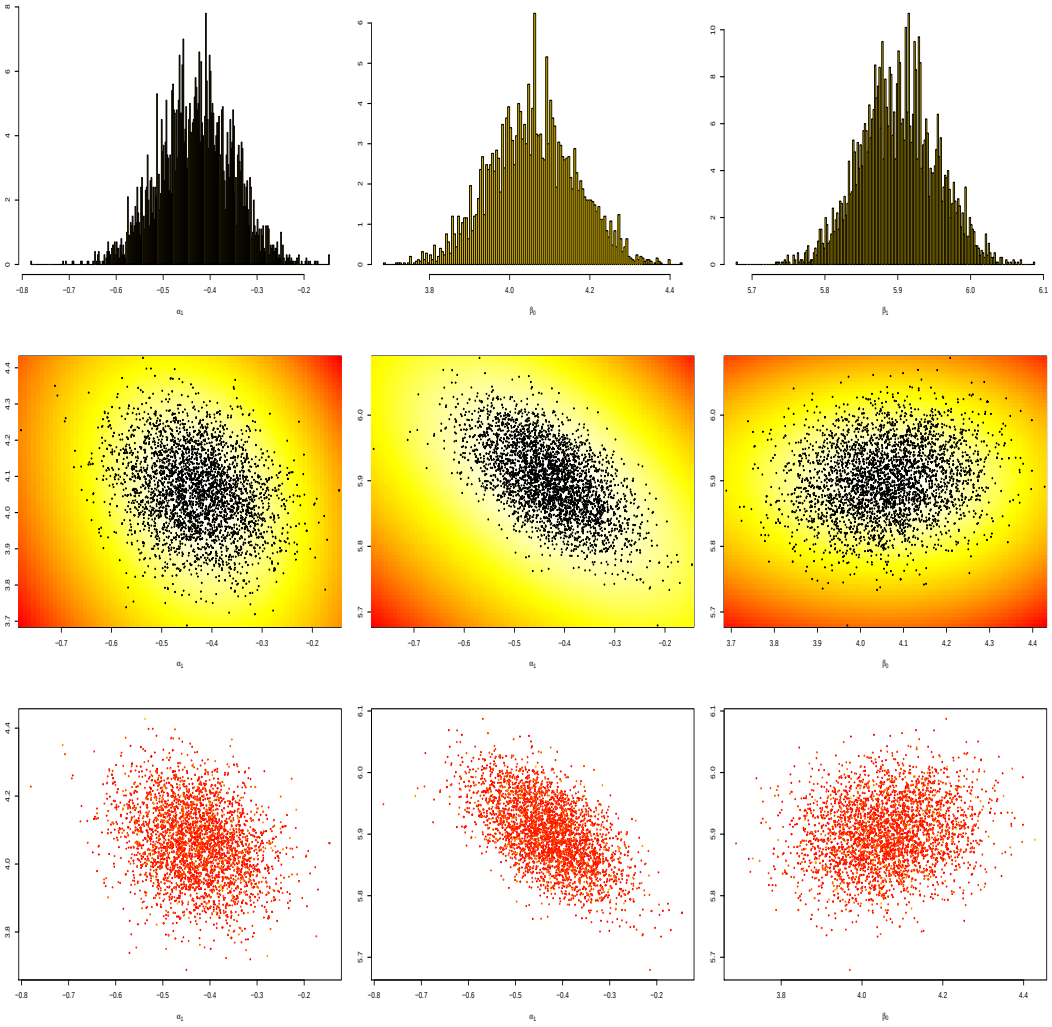

Figure 3: Repartition of $5,000$ resampled points after 5 iterations of the Rao-Blackwellized D-kernel PMC sampler for the contingency table example: *first row:* histograms of the components α_1, β_0 and β_1, *second row:* scatterplot of the points (α_1, β_0), (α_1, β_1), and (β_0, β_1), on the profile slices of the loglikelihood, *third row:* scatterplot of the same points with a color representation of the corresponding likelihoods: the darker hues are for higher likelihoods and the range is set by the original distribution of the likelihood (represented in Figure 5 for the first four iterations).

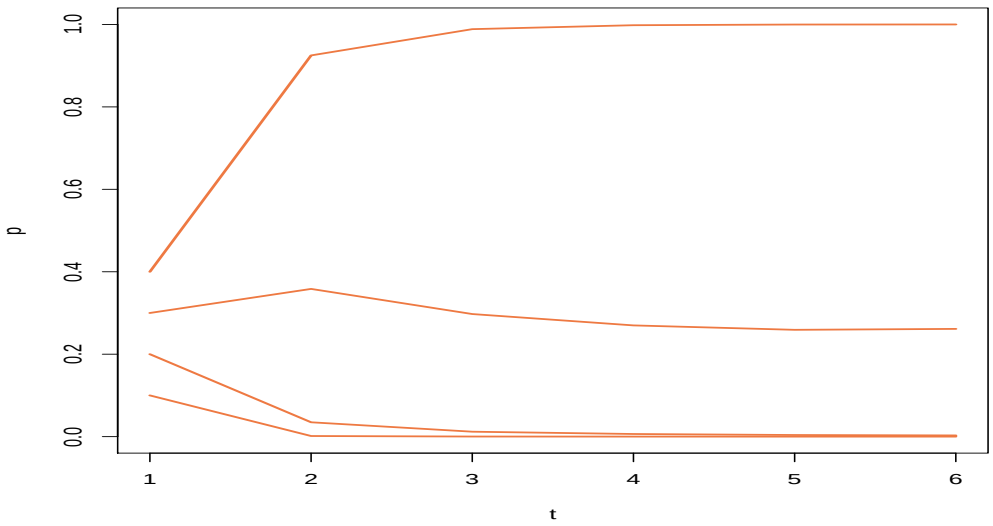

Figure 4: Convergence of the cumulated weights of the 5th, 6th, 7th and 8th components of the mixture for the contingency table example.

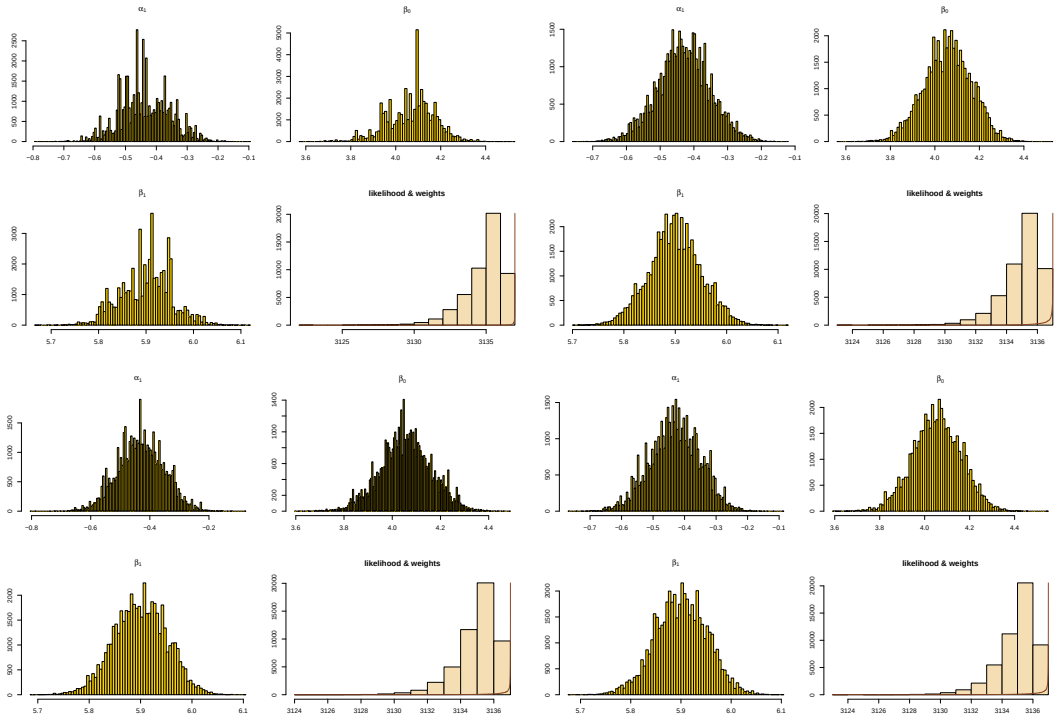

Figure 5: Evolution of the samples over 4 iterations of the Rao-Blackwellized D-kernel PMC sampler for the contingency table example *(the output from each iteration is a vignette of four graph, to be read from left to right and from top to bottom)*: histograms of the resampled samples of α_1, β_0 and β_1 of size $50,000$ and *(lower right of each vignette)* loglikelihood and the empirical cdf of the importance weights.

Csizàr, I. and Tusnàdy, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions*, pages 205–237. Suplementary Issue Number 1.

Del Moral, P. (2004). *Feynman-Kac formulae*. Probability and its Applications. Springer-Verlag, New York.

Douc, R. and Moulines, E. (2005). Limit theorems for properly weighted samples with applications to sequential Monte Carlo. Technical report, TSI, Telecom Paris.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Gilks, W., Roberts, G., and Sahu, S. (1998). Adaptive Markov chain Monte Carlo. *J. American Statist. Assoc.*, 93:1045–1054.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3):375–395.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hesterberg, T. (1998). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37:185–194.

Iba, Y. (2000). Population-based Monte Carlo algorithms. *Trans. Japanese Soc. Artificial Intell.*, 16(2):279–286.

Künsch, H. (2004). Recursive Monte Carlo filters: algorithms and theoretical analysis. *Ann. Statist.* (to appear).

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121.

Ripley, B. (1987). *Stochastic Simulation*. J. Wiley, New York.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 2 edition.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120.

Rubin, D. (1987). A noniterative sampling importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. (In the discussion of Tanner and Wong paper). *J. American Statist. Assoc.*, 82:543–546.

Rubin, D. (1988). Using the SIR algorithm to simulate posterior distributions. In Bernardo, J., Degroot, M., Lindley, D., and Smith, A., editors, *Bayesian Statistics 3: Proceedings of the Third Valencia International Meeting, June 1-5, 1987*. Clarendon Press.

Rubinstein, R. (1981). *Simulation and the Monte Carlo Method*. J. Wiley, New York.

Sahu, S. and Zhigljavsky, A. (1998). Adaptation for self regenerative MCMC. Technical report, Univ. of Wales, Cardiff.

Sahu, S. and Zhigljavsky, A. (2003). Self regenerative Markov chain Monte Carlo with adaptation. *Bernoulli*, 9:395–422.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, 22:1701–1786.

A Convergence theorem for triangular arrays of random variables

In this section, we recall some convergence results for triangular arrays of random variables (see Cappé et al., 2005 or Douc and Moulines, 2005 for more details, including the proofs). We will use these results to study the asymptotic behavior of the PMC algorithm. In the following, let $\{U_{N,i}\}_{N\geq 1, 1\leq i\leq N}$ be a triangular array of random variables defined on the same measurable space $(\Omega, \mathcal{A})$, let $\{\mathcal{G}_N\}_{N\geq 1}$ be a sequence of σ-algebras included in $\mathcal{A}$, the symbol $X_N \longrightarrow_{\mathbb{P}} a$ means X_N converges in probability to a as N goes to infinity.

The definitions and theorems we need in the above proofs are given below.

Definition A.1. *We say that $\{U_{N,i}\}_{N\geq 1, 1\leq i\leq N}$ is independent given $\{\mathcal{G}_N\}_{N\geq 1}$ if, $\forall N \geq 1$, the random variables $U_{N,1}, \ldots, U_{N,N}$ are independent given $\mathcal{G}_N$.*

Definition A.2. *A sequence of random variables $\{Z_N\}_{N\geq 1}$ is said to be bounded in probability if*

$$\lim_{C\to\infty} \sup_{N\geq 1} \mathbb{P}\left[|Z_N| \geq C\right] = 0.$$

Theorem A.1. *If*

(i) $\{U_{N,i}\}_{N\geq 1, 1\leq i\leq N}$ *is independent given* $\{\mathcal{G}_N\}_{N\geq 1}$;

(ii) the sequence $\left\{\sum_{i=1}^{N} \mathbb{E}[|U_{N,i}||\mathcal{G}_N]\right\}_{N\geq 1}$ *is bounded in probability* ;

(iii) $\forall \eta > 0$, $\sum_{i=1}^{N} \mathbb{E}[|U_{N,i}|\mathbb{I}_{|U_{N,i}|>\eta}|\mathcal{G}_N] \longrightarrow_{\mathbb{P}} 0$;

then $\sum_{i=1}^{N} (U_{N,i} - \mathbb{E}[U_{N,i}|\mathcal{G}_N]) \longrightarrow_{\mathbb{P}} 0.$

Theorem A.2. *If*

(i) $\{U_{N,i}\}_{N\geq 1, 1\leq i\leq N}$ *is independent given* $\{\mathcal{G}_N\}_{N\geq 1}$;

(ii) $\forall N \geq 1, \forall i \in \{1, \ldots, N\}$, $\mathbb{E}[|U_{N,i}||\mathcal{G}_N] < \infty$;

(iii) $\exists \sigma^2 > 0$ *such that* $\sum_{i=1}^{N} \left(\mathbb{E}[U_{N,i}^2|\mathcal{G}_N] - (\mathbb{E}[U_{N,i}|\mathcal{G}_N])^2\right) \longrightarrow_{\mathbb{P}} \sigma^2$;

(iv) $\forall \eta > 0$, $\sum_{i=1}^{N} \mathbb{E}[U_{N,i}^2\mathbb{I}_{|U_{N,i}|>\eta}|\mathcal{G}_N] \longrightarrow_{\mathbb{P}} 0$;

then,

$$\forall u \in \mathbb{R}, \quad \mathbb{E}\left[\exp\left(iu\sum_{i=1}^{N}(U_{N,i} - \mathbb{E}[U_{N,i}|\mathcal{G}_N])\right)\middle|\mathcal{G}_N\right] \longrightarrow_{\mathbb{P}} \exp\left(-\frac{u^2\sigma^2}{2}\right).$$

Unité de recherche INRIA Futurs
Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399